LA-UR-22-21008

Approved for public release; distribution is unlimited.

Title:	Using Deep Mutational Data and Machine Learning to Guide Outbreak and Pandemic Response
Author(s):	Hu, Bin Gans, Jason David Lin, Youzuo Li, Po-E Chain, Patrick Sam Guy
Intended for:	Report
Issued:	2022-02-07

Using Deep Mutational Data and Machine Learning to Guide Outbreak and Pandemic Response

Bin Hu, Jason Gans, Po-E Li, Youzuo Lin, and Patrick Chain
Los Alamos National Laboratory
Contact: bhu@lanl.gov

A significant fraction of pathogens known to infect humans originate in non-human (zoonotic) hosts (Taylor, Latham, and Woolhouse 2001), and new and emerging pathogens continue to spill over into the human population more frequently at an alarming rate (e.g., SARS, MERS, Cholera, etc.). The recent outbreaks of Ebola virus in West Africa and the ongoing SARS-CoV-2 pandemic demonstrate the need for rapid and reliable assessments of *viral phenotype* information to help inform scientists and policy makers how best to control the spread of disease. Further understanding of the virus pathogenic evolutionary space and potential trajectory could guide appropriate control measures to limit the spread of a new virus throughout the local and global human population.

Once a viral disease begins to circulate, reliable diagnostics, protective vaccines and therapeutic antibodies are essential tools for preventing, monitoring, and managing disease spread. However, the efficacy of these tools can be diminished by mutations in viral genomes (as has been observed in the ongoing pandemic), and the delay between the emergence of new viral strains and redesign of vaccines and diagnostics allows for continued viral transmission. Given the combinatorial explosion of potential mutations that could enable a virus to "escape" diagnostics, vaccines and antibodies, and the high cost of biomedical research, it is essential to focus countermeasure development efforts only on viral strains that pose the highest risk to society. Towards this end, the questions we ask are: Is it possible to predict the most likely evolutionary trajectory of circulating genomes and anticipate novel variants before they emerge? Is it possible to assess the risk of future variants by computationally predicting key virulence determinants and exploring the evolutionary space for pathogenicity?

The ability to answer these questions hinges on our ability to predict viral phenotype (e.g., the physical properties of a virus) from viral genotype (e.g., the genome sequence of a virus). Understanding the dependence of phenotype on genotype is a longstanding, "grand challenge" problem in biology. Thanks to recent progress in machine learning (ML), protein structure prediction and the accumulation of large biological datasets, the field of genotype to phenotype prediction appears poised for rapid progress. In particular, the large number of SARS-CoV-2 genome sequences ($> 10^7$) provides an unprecedented catalog of viral genotypes. While many challenges will need to be overcome, the DOE national laboratories are well positioned to make significant contributions to solving the problem of predicting phenotype from genotype.

We propose to utilize ML and high-throughput mutational data to solve the problem. As a pilot project to test the feasibility of this approach, we developed three neural network models with a training data set generated from a deep mutational scanning (DMS) library. DMS is a recently developed high-throughput technology that can generate $>10^5$ random mutations (Fowler and Fields 2014). It has been applied to

SARS-CoV-2 receptor binding domain (RBD) to generate 116,257 unique mutated RBD sequences with linked expression levels and binding kinetics to the human receptor protein of the virus (ACE2) (Starr et al. 2020). Follow up studies have also measured the neutralization antibody (nAb) binding kinetics of the mutated sequences (Greaney et al. 2021).

We have developed a one-hot encoded, densely connected NN as the proof-of-concept test. This simple model was able to predict fairly accurately the RBD expression and binding to ACE2 ($R2 = 0.76$). It only takes less than a second for the model to predict the effect of an arbitrary mutation, regardless of the combinatorial complexity, on a consumer PC. Major limitations of the current feature set in such a densely connected NN includes lack of biochemical property of RBD, including charge of the amino acids, sequence, and 3-D structure of the protein and many others. Recently, principal component analysis of amino acid biochemical properties and graph neural network (GNN) to learn protein properties have been combined to predict antibody binding and enzyme activities in five proteins (Gelman et al. 2021). Using a similar approach, we have developed a GNN model to study RBD and are currently evaluating the model. This model may also provide a mechanism to explore the RBD pathogenic evolutionary space computationally.

Combining DMS and deep learning, we can predict the mutational effects of SARS-CoV-2 RBD, both in its expression and binding to the ACE2 receptor. We have also identified methods to accommodate the biochemical properties of amino acids, sequence, and 3-D protein structural information. We are working on evaluating different ML models and will deploy either the best model or an ensemble of models to our existing SARS-CoV-2 sequence monitoring workflow, which synchronizes all the submitted SARS-CoV-2 sequences from GISAID and NIH daily. The upgraded workflow will be able to rank the latest mutations based on potential threat level.

References

Fowler, Douglas M., and Stanley Fields. 2014. “Deep Mutational Scanning: A New Style of Protein Science.” *Nature Methods* 11 (8): 801–7.

Gelman, Sam, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero, and Anthony Gitter. 2021. “Neural Networks to Learn Protein Sequence–Function Relationships from Deep Mutational Scanning Data.” *Proceedings of the National Academy of Sciences* 118 (48).

Greaney, Allison J., Tyler N. Starr, Pavlo Gilchuk, Seth J. Zost, Elad Binshtein, Andrea N. Loes, Sarah K. Hilton, et al. 2021. “Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain That Escape Antibody Recognition.” *Cell Host & Microbe* 29 (1): 44-57.e9.

Starr, Tyler N., Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, et al. 2020. “Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.” *Cell* 182 (5): 1295-1310.e20.

Taylor, L. H., S. M. Latham, and M. E. Woolhouse. 2001. “Risk Factors for Human Disease Emergence.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 356 (1411): 983–89.